全国第七届研究生数学建模竞赛

题 目 基因表达谱的复杂网络研究

摘 要：

本文采用复杂网络理论，整合基因关联信息和文献中的结果，得到一些关于结肠癌标志基因的可能的结果。首先利用分类信息指数对数据进行初步筛选，选出了 314 个基因。对选出的基因分别做肿瘤样本和正常样本的相关系数矩阵，利用 Kruskal 算法分别对两个相关系数矩阵做最小生成树，然后通过两种方法比较选出阈值，建立起节点间的连边关系，得到致病前后的两个网络。根据复杂网络中的相关理论，分别对肿瘤样本和正常样本进行社区划分，最后通过观察两个样本的网络系统，分析致病前后基因的变化情况，建议了结肠癌的特征基因。

关键字：相关系数矩阵 最小生成树 复杂网络 社区结构

参赛队 k0000057

队员姓名 齐景超 张东 张珍

参赛密码 _____

（由组委会填写）

中山大学承办

一 问题的重述

癌症起源于正常组织在物理或化学致癌物的诱导下，基因组发生的突变，即基因在结构上发生碱基对的组成或排列顺序的改变，因而改变了基因原来的正常分布(即所包含基因的种类和各类基因以该基因转录的mRNA的多少来衡量的表达水平)。所以探讨基因分布的改变与癌症发生之间的关系具有深远的意义。

DNA微阵列（DNA microarray），也叫基因芯片，是最近数年发展起来的一种能快速、高效检测DNA片段序列、基因表达水平的新技术。它将数目从几百个到上百万个不等的称之为探针的核苷酸序列固定在小的（约$1\,cm^2$）玻璃或硅片等固体基片或膜上，该固定有探针的基片就称之为DNA微阵列。根据核苷酸分子在形成双链时遵循碱基互补原则，就可以检测出样本中与探针阵列中互补的核苷酸片段，从而得到样本中关于基因表达的信息，这就是基因表达谱，因此基因表达谱可以用一个矩阵或一个向量来表示，矩阵或向量元素的数值大小即该基因的表达水平。

随着大规模基因表达谱（ Gene expression profile， 或称为基因表达分布图） 技术的发展，样本类各种组织的正常的基因表达已经获得，各类病样本的基因表达分布图都有了参考的基准，因此基因表达数据的分析与建模已经成为生物信息学研究领域中的重要课题。如果可以在分子水平上利用基因表达分布图准确地进行肿瘤亚型的识别，对诊断和治疗肿瘤具有重要意义。因为每一种肿瘤都有其基因的特征表达谱。从DNA 芯片所测量的成千上万个基因中，找出决定样本类别的一组基因“标签”，即“信息基因” （informative genes ）是正确识别肿瘤类型、给出可靠诊断和简化实验分析的关键所在，同时也为抗癌药物的研制提供了捷径。

通常由于基因数目很大，在判断肿瘤基因标签的过程中，需要剔除掉大量“无关基因”，从而大大缩小需要搜索的致癌基因范围。事实上，在基因表达谱中，一些基因的表达水平在所有样本中都非常接近。例如，不少基因在急性白血病亚型（ALL,AML）两个类别中的分布无论其均值还是方差均无明显差别，可以认为这些基因与样本类别无关，没有对样本类型的判别提供有用信息，反而增加信息基因搜索的计算复杂度。因此，必须对这些“无关基因” 进行剔除。经过10余年的努力，在基因表达谱分析方面取得了长足的进展，但是仍然有很多基本的问题没有解决，主要有下边几个方面，

（1） 由于基因表示之间存在着很强的相关性，所以对于某种特定的肿瘤，似乎会有大量的基因都与该肿瘤类型识别相关，但一般认为与一种肿瘤直接相关的突变基因数目很少。如何从上述观点出发，选择最好的分类因素？

（2） 相对于基因数目，样本往往很小，如果直接用于分类会造成小样本的学习问题，如何减少用于分类识别的基因特征是分类问题的核心，事实上只有当这种特征较少时，分类的效果才更好些。也就是如何从分类的角度确定相应的基因“标签”？

（3） 基因表达谱中不可避免地含有噪声，有的噪声强度甚至较大，对含有噪声的基因表达谱提取信息时会产生偏差。通过建立噪声模型，分析给定数据中的噪声能否对确定基因标签产生有利的影响？

（4） 在肿瘤研究领域通常会已知若干个信息基因与某种癌症的关系密切，建立融入了这些有助于诊断肿瘤信息的确定基因“标签”的数学模型。比如临

床有下面的生理学信息：大约 90%结肠癌在早期有 5 号染色体长臂 APC 基因的失活，而只有 40%~50%的 ras 相关基因突变。

（5） 从系统生物学的角度出发，整合基因组、蛋白质组、代谢组、以及临床等各种数据，找到癌症不同发展时期的标志性基因。

本文采用复杂网络理论，从基因表达谱数据构建基因之间的复杂网络关系。通过分析正常样本和肿瘤样本网络之间的结构差异，以及文献中发现的结肠癌相关基因在网络上的分布特征，试图评价文献中建议的有关基因，寻找出结肠癌的标志性基因。

二 问题分析

癌症起源于正常组织在物理或化学致癌物的诱导下，基因组发生的突变，即基因在结构上发生碱基对的组成或排列顺序的改变，因而改变了基因原来的正常分布（即所包含基因的种类和各类基因以该基因转录的 mRNA 的多少来衡量的表达水平）。所以探讨基因分布的改变与癌症发生之间的关系具有深远的意义。肿瘤的发生发展似乎不是相关基因发生遗传改变后，简单的作用叠加结果，而是一种细胞生长、分化异常的分子网络病。一种肿瘤有多个基因参与，一个基因参与多种肿瘤，任何一个基因都不是独立执行功能，而是作为细胞网络中的一个环节，与其他基因相互协调来完成一定的生物学过程[1]。从系统的角度去观察和分析肿瘤的生物学问题，是生物信息学对肿瘤研究的一个新的方向。现在肿瘤基因表达谱分析从研究单一肿瘤特征基因转向研究肿瘤基因表达调控网络，以期实现对基因功能的整体认识和把握。基因表达调控网络的研究对于寻找和识别样本类致病基因具有特别重要的意义。

复杂网络理论是近年来发展起来的一个重要的交叉。对于一个复杂的系统，很多时候我们不能够单独通过分析系统内元组来反应系统性质。复杂系统是由微观层次上的海量个体所组成，个体之间存在着作用。把个体抽象为网络节点，而个体之间的相互作用抽象为节点之间的边，则复杂系统就可以用一个复杂网络来描述。

本文主要通过多个序列构造复杂网络的方法研究基因之间的关系。我们的想法是结合已给的肿瘤样本(cancer)和正常样本(normal)的基因表达谱的数据，通过分析基因间的关系以及肿瘤样本(cancer)和正常样本(normal)的基因变化信息，从这两者出发来分别生成关于肿瘤样本(cancer)和正常样本(normal)基因谱表达相关的复杂网络，通过研究网络的相关性质来揭示基因与疾病发生的一些内在关系。

三 数据

3.1 基因表达谱

本文的实验数据集包含 22 个正常组织样本和 40 个结肠癌组织样本，每个样本包含 2000 个基因的表达数据。首先对样本数据进行归一化，另外，数据的特征维数 2000，远远高于样本个数 62。因此，有必要对数据进行过滤和降维。测量的数千个基因的表达水平有的差异很大，只有少部分基因同样本的类别有很强的相关性，而大部分基因与样本的类别不相关，对分类没有什么贡献，这些基因也应该从数据中滤除。 考虑到这些问题，我们采用了分类信息指数方法(information index to classification, ⅡC)[2]，作为衡量尺度来挑选每一类的“主基因”，即选取每一类中具有信息分类指数最高值的基因作为类的“主基因”，再将每一类的主基因作为特征基因来建立分类模型。分类信息指数公式为：

$$IIC(i)=\frac{1}{2}\frac{|\mu_1(i)-\mu_2(i)|}{(\sigma_1(i)+\sigma_2(i))}+\frac{1}{2}\ln\left(\frac{\sigma_1(i)^2+\sigma_2(i)^2}{2\sigma_1(i)\sigma_2(i)}\right) \qquad (1)$$

其中，$\mu_1(i)$，$\mu_2(i)$分别表示第i个基因在正常组织样本和结肠癌组织样本中的中表达水平的均值；$\sigma_1^{\,2}(i)$，$\sigma_2^{\,2}(i)$分别为该基因表达水平的标准差。

⑴式中的第一项是 Golub 等样本定义的“ 信噪比”指标；第二项体现了表达水平分布方差的不同对样本分类的贡献。依据该指标 ，即使基因在两类不同样本中表达水平的均值相同 ，只要分布方差出现大的差别 ，仍然可以获得较大的分类信息指数。

根据上式计算结肠癌基因表达数据中的 2000 个基因的分类信息指数，大部分基因的分类信息指数在 0 到 0.2 之间，仅有少部分基因的大于 0.2（如图 1）。保留指数大于 0.2 的 314 个基因用于下一步的分析，这样就大大缩小了基因选择的特征空间，降低了数据维数。通过对数据的初步筛选，剔除掉大量“无关基因”，大大缩小需要搜索的致癌基因范围。

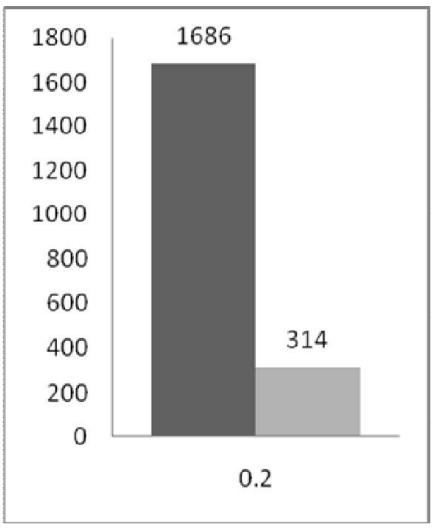

图 1 结肠癌基因表达谱中 2000 个基因的分类信息指数分布。

3.2 文献中发现的结肠癌相关基因

另外在撰写本文的准备过程中，我们查阅了大量的有关文献。与已有文献的结果进行比较 ，发现所选特征基因中包含了一些已被实验证实的与癌症相关的重要基因 ，这些基因在癌症基因调控网络中起关键作用，例如，Guyon 等样本以 62 个样本为学习样本，利用线性支持向量机找出了 7 个基因：H64807、T62947、R88740、H81558、T94579、M59040 和 H08393[3]，刘全金等样本以 SVM 的灵敏度分析方法，仍以 62 个样本为学习样本，选取出的 7 个基因分别为 H08393，H20709，M82919，T51849，T57619，K02268，R88740[4]，刘全金等样本以浮动顺序搜索算法得到 M76378 和 U19969 这 2 个基因在肿瘤组织样本中呈下调表达 ，而在正常组织样本中相对上调表达；J05032 在肿瘤组织样本中为上调表达 ，在正常组织样本中又相对下调表达[5]，Xiaosheng Wang and Osamu Gotoh 用软件算方法找出了 19 个基因：M63391, M76378, J02854, M26383, T60155, M22382, X12671, T96873, X86693, J05032, U25138, T60778, M91463, R87126, T51571, T92451, U09564, R97912, L41559[6]，Xue Wu Zhang 等样本利用独立成分分析方法找出 10 个基因：H06524，J02854，H43887，L05144，M36634，M27190，R54097，J05032，X62048，M26383[7]，李建更等样本，引入一种最高得分对(TSP)方法 ，处理一组包含 40 个肿瘤和 22 个正常样本的结肠癌微阵列数据 ，得到标志基因对(M36634，J05032)并构建双基因分类器，两基因在正常和肿瘤样本中的起峰早晚恰好相反[8]，张娅等样本利用 kmeans IIC 法得出 T49941 所对应的基因可以用

于结肠癌的风险预测诊断[9]。综上除去各文献间重复基因，一共得到了 40 个基因（如表 1）。我们要探寻的结肠癌的特征基因极有可能包含在这 40 个基因中，这对我们后续的研究具有重要的参考价值。

在这 40 个基因中，其中 6 个基因在我们根据分类信息指数值对数据进行筛选的过程中被剔除了。所以我们选择剩下的 34 个基因作为我们研究的参考。如表 1 所示。表 2 为 34 个基因在已给数据库和我们的网络中的变化对应表。

表 1 文献中发现的 34 个结肠癌相关基因。Gene ID 是基因表达谱中的基因编号。

Gene ID	GenBank Acc. No	Mapped region
14	H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN);.
245	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
249	M63391	Human desmin gene, complete cds.
415	T60155	ACTIN, AORTIC SMOOTH MUSCLE (HUMAN);.
493	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
513	M22382	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN);.
581	T51571	P24480 CALGIZZARIN.
625	X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
792	R88740	ATP SYNTHASE COUPLING FACTOR 6, MITOCHONDRIAL PRECURSOR (HUMAN);.
822	T92451	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN);.
897	H43887	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
1060	U09564	Human serine kinase mRNA, complete cds.
1115	R97912	SERINE/THREONINE-PROTEIN KINASE IPL1 (Saccharomyces cerevisiae)
1227	T96873	HYPOTHETICAL PROTEIN IN TRPE 3'REGION (Spirochaeta aurantia)
1346	T62947	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)
1387	L05144	PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC (HUMAN);contains Alu repetitive element;contains element PTR5 repetitive element ;.
1400	M59040	Human cell adhesion molecule (CD44) mRNA, complete cds.
1423	J02854	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element TAR1 repetitive element ;.
1472	L41559	Homo sapiens pterin-4a-carbinolamine dehydratase (PCBD) mRNA, complete cds.
1473	R54097	TRANSLATIONAL INITIATION FACTOR 2 BETA SUBUNIT (HUMAN);.
1494	X86693	H.sapiens mRNA for hevin like protein.

1570	H81558	PROCYCLIC FORM SPECIFIC POLYPEPTIDE B1-ALPHA PRECURSOR (Trypanosoma brucei brucei)
1635	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
1668	M82919	Human gamma amino butyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds.
1671	M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds.
1771	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds.
1772	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
1843	H06524	GELSOLIN PRECURSOR, PLASMA (HUMAN);.
1892	U25138	Human MaxiK potassium channel beta subunit mRNA, completecds.
1897	U19969	Human two-handed zinc finger protein ZEB mRNA, partial cds.
1917	M91463	Human glucose transporter (GLUT4) gene, complete cds.
1924	H64807	PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
1935	X62048	H.sapiens Wee1 hu gene.
1967	T60778	MATRIX GLA-PROTEIN PRECURSOR (Rattus norvegicus)

四 方法

本文从基因表达谱建立基因之间的关系，得到基因关系网络。通过整合基因关系网络和文献中发现的结肠癌相关基因，对文献中的结果进行评价，并且试图发现结肠癌标志性基因。

本文的方法是首先对结肠癌样本和正常样本的基因表达的变化比较明显的进行筛选分析，表中提供 40 个结肠癌样本和 22 个正常样本的 2000 个基因表达数据，首先从序列中筛选 314 个变化比较明显的基因进行问题的研究。

然后分别计算结肠癌样本(cancer)和正常样本(normal)各个基因间的相似性，得到相似矩阵。分析这些基因点的联系，选择一个相似性的阈值来分别建立复杂网络，用邻接矩阵表示。(如果相似性大于该阈值的则这两个点相连接，在邻阶矩阵中用 1 表示；反之，如果相似性小于于该阈值的则这两个点不连接，在邻阶矩阵中用 0 表示)。其中关键的步骤是阈值的选取。本文提出的解决策略是，从关联系数矩阵得到最小生成树作为基因之间关系的骨架，然后再把文献中发现的相关基因之间的关系考虑进来，得到客观的阈值。

4.1 距离矩阵

我们考查结肠癌基因表达数据中筛选出来的 314 个变化比较明显的基因，表示为，

$$T^0 = \left\{ T^0_{m,n}, m = 1,2,\cdots,M; n = 1,2,\cdots,N \right\}, \quad (2)$$

其中 $T^0_{m,n}$ 是第 n 个基因在第 m 个样本的基因数据，其中 $N = 2000$， M 是样本个

数。相关系数矩阵为 R：

$$R_{n_1,n_2}=\frac{\sum_{m=1}^{M}\left[T_{n_1,m}-\left(T_{n_1,m}\right)\right]\left[T_{n_2,m}-\left(T_{n_2,m}\right)\right]}{\sqrt{\left[T_{n_1,m}-\left(T_{n_1,m}\right)\right]^2\left[T_{n_2,m}-\left(T_{n_2,m}\right)\right]^2}} \quad (3)$$

那么基因间的欧几里得距离就可以用以下定义的距离矩阵 D 定量描述：

$$D_{n_1,n_2}=1-\left|R_{n_1,n_2}\right| \quad (4)$$

4.2 最小生成树

最小生成树是图论中的基本概念。我们从距离矩阵中抽取出最小生成树，用 $N-1$ 条边连接所有基因节点，形成一个无圈图。在形成的最小生成树中，要保证所有基因间的距离之和最小，也即相关系数之和最大，且是无圈图。那么，基因间的其它关系就被过滤掉了。原则上来讲，真正直接相关的基因之间的关联系数最大，因此可以认为最小生成树保留了基因之间的真正关系。因为一个基因可以和多个基因直接相关，所以很多的关系被丢掉。丢掉的关系将在后边的步骤中被找回。

我们采用 Kruskal 算法来生成最小生成树：

i) 每一个基因是一个初始孤立点；

ii) 找到距离最近的两个基因，并将其连接，作为生成树；

iii) 在剩下的基因中找到与生成树上的基因距离最短，且连接以后不产生圈的，其与生成树相应的基因连接，作为新的生成树；

iv) 重复第 iii)步，直到把所有的基因连接完。

我们用筛选后的 314 个基因数据（我们对这 314 个基因重新做了编号，其与原数据库中的编号的对应表见附表），对结肠癌样本、正常样本分别用两种方法得到了最小生成树。两个最小生成树的节点也即基因，一定是相同的，且都有 314 个节点，313 条边。图 2 给出了正常样本中得到的最小生成树。

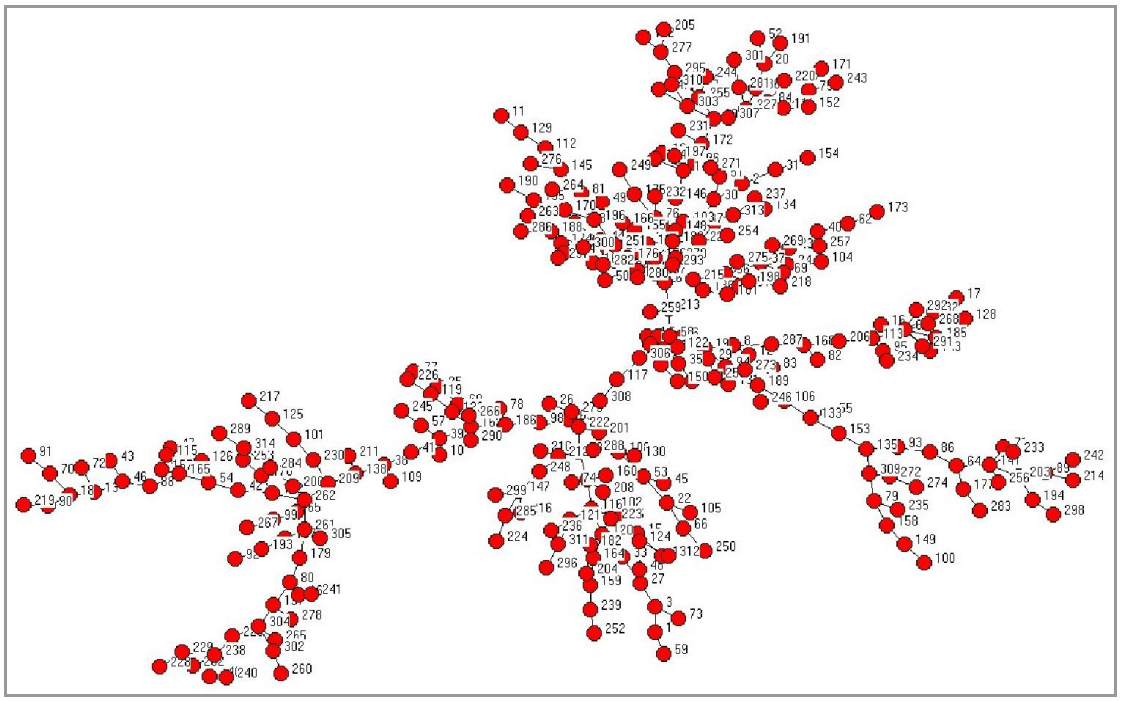

图 2 正常样本数据的最小生成树

4.3 基因关系网络

4.3.1 阈值的选取

如前所述，最小生成树给出了基因之间的部分连接，但是很多基因之间的关系被丢掉。另一方面，文献中发现的结肠癌相关基因，为我们提供了重要的参考信息，但是这些信息包含着很大的偶然性，也就是噪声。在此我们将把这两部分信息整合在一起，得到一个客观的构建基因关系网络的阈值。

我们首先抽取出如图 2 所示的生成树。它给我们提供了高可信度的链接，不足之处是包含的信息不够多，一些重要的关系被忽略了。我们再根据得病前后两类样本信息变化。然而，这里也可能产生噪声边。

为了消除这样的噪声，同时恢复我们需要的链接，我们在最小生成树上标注文献中发现的结肠癌相关基因（或者从分类信息筛选出来的可信的基因）。这些被标注的基因如果在最小生成树上相连，则记录他们之间的距离。找到这些距离中最大的一个，这个值可以作为构造基因关系网络的阈值。

分类信息指数最大的 50 个节点肯定在最小生成树上，它们之间可能直接相连，也可能彼此没有直接相连。阈值的选取就是找到这 50 个节点在最小生成树中的位置，计算直接相连的节点间的距离。在这个过程中，我们选取最大的那个作为阈值，在肿瘤样本生成树中记为 $DDI_{\min}$，在正常样本的生成树中记为 $DNI_{\min}$。换言之，找到这 50 个节点在树中彼此直接相连的节点间的相关系数，选择这些基因中相关系数最小的最为阈值。通过这个距离阈值（最小的相关系数），我们就得到所需要的复杂网络。

阈值的另外一种常用的选取方法，与上面类似，也是从上面得到最小生成树出发。整合相关文献中已知的肿瘤致病基因，我们收集到 34 个这样的基因。用这 34 个基因重复上面的过程，得到阈值，肿瘤样本的记为 $DDII_{\min}$，在正常样本

的生成树中记为 $DNII_{\min}$ 建立网络。

通过两种方法的比较，我们发现阈值的计算是很接近的。分别为：

$$\begin{cases} DNI = 0.7032 \\ DDI = 0.6378 \end{cases}, \quad \begin{cases} DNII = 0.6995 \\ DDII = 0.6239 \end{cases}$$

我们选取 $DDII_{\min}$，$DNII_{\min}$ 作为阈值，来建立网络。这样在一定程度上减少了一些噪声边的产生，避免了偶然因素可能引起的阈值选取的不稳定性，同时也恢复了我们需要的连接。

4.3.2 网络的生成

肿瘤样本网络以及正常样本网络的阈值选定后，利用我们在数据处理中选定的 314 个基因建立网络。以肿瘤样本网络为例，先算出肿瘤样本中这 314 个基因的相关系数矩阵。当任意两个基因的相关系数大于阈值 0.6239（$DDII = 0.6239$）时，我们就认为这两个基因是有相互作用的，在它们之间画一条边；当任意两个基因的相关系数小于阈值 0.6239（$DDII = 0.6239$）时，我们就认为这两个基因是没有相互作用的，它们之间就没有直接的边相连。这样我们就得到了肿瘤样本的基因相互作用网络。在相关系数矩阵中，把大于 0.6239 的值改为 1，小于 0.6239 的改为 0，主对角线上元素设为 0，这样就由相关系数矩阵得到了邻接矩阵 MD。邻接矩阵中的 1 就表示网络中有连边；邻接矩阵中的 0 就表示网络中没有连边。

利用同样的方法，以阈值 $DNII = 0.6995$，可以建立正常样本基因间的网络以及邻接矩阵 MN。

4.4 社区的划分

复杂网络的结构是不均匀的，往往存在很多连接致密的集团，在这些集团之间只有很少边形成的松散的连接。这些致密的结构往往与功能有着密切的关系，因此受到普遍的关注。当前普遍采用的划分社区的方法是 Newman-Girvan 算法。这一算法中定义了一网络描述社区性质的参量 Q。Q 值越大则表明社区性质越好，因此一般把 Q 的极大值对应的划分作为社区划分结果。

参量 Q 的定义为：

$$Q = \sum_i (e_{ii} - a_i^2) = TrE - \left\| E^2 \right\|$$

其中 E 是 n 阶对称矩阵，元素 e_{ij} 是网络中所有边中连接社团（模块） 的点 i 和社团的点 j 的那部分边。矩阵的迹 $TrE = \sum_i e_{ii}$ 是网络中连接同一社团的点的那部分边，而行和（或列和）$a_i = \sum_j e_{ij}$ 给出连接社团 i 的点的那部分边。如果网络中两点之间有一条边的概率是相等的，不管最后是否属于同一个社团，将有 $e_{ij} = a_{ij}$。$\left\| E \right\|$ 表示矩阵 E^2 的元素之和。该参量度量了两点之间有一条边的概率与这两点属于同一社团的实际情况的相关程度。

社区划分反映基因间的功能关系，而在网络模块中，可以发现网络发生了明显的改变。首先我们画出正常样本网络，用 Newman-Girvan 的划分算法对得到的网络进行分块。当把正常样本网分成 14 个社区时，得到的聚类系数最大，为 Q=0.596（如表 3），这样就把网络分成了 14 个大的功能模块。如图 3 所示，即为正常样本网络的社区结构（每种颜色代表一个社区）。可以看出，各个社区结构中的节点数目分布并不均匀，并且存在很多孤立节点。社区内节点间的连接比较紧密，而不同社区间的连接比较稀疏。

表 3 正常样本网络不同社区划分的Q值

社区数	10	11	12	13	14	15	16
Q 值	0.362	0.366	0.450	0.594	0.596	0.596	0.593

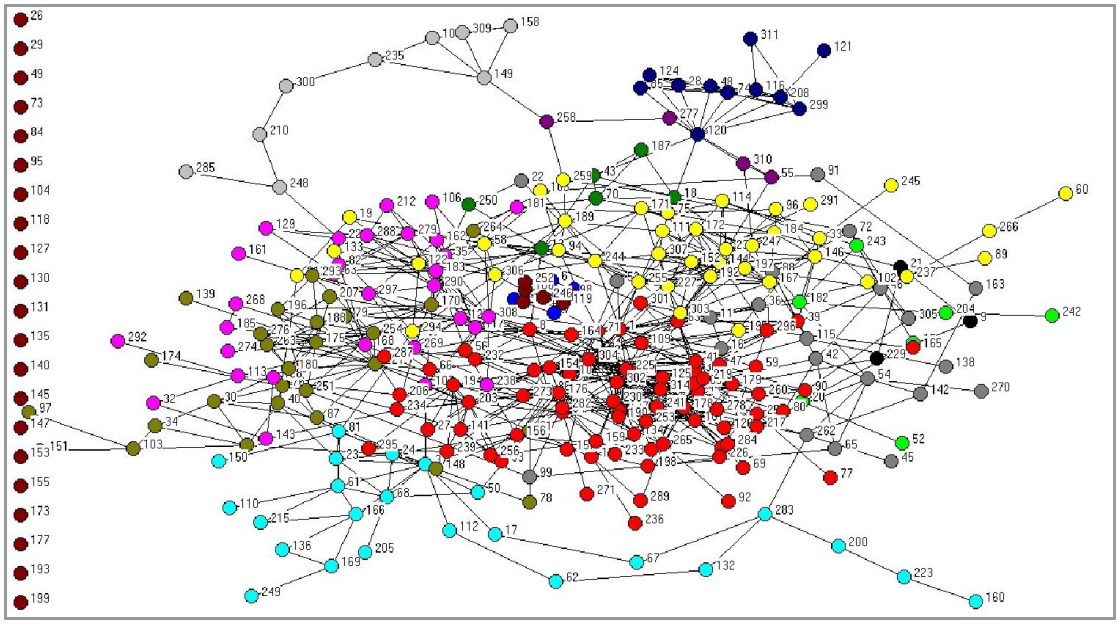

图 3　正常样本网络社区结构图，不同的颜色表示不同的社区

同样用 Newman-Girvan 的划分算法，我们画出肿瘤样本的网络，把肿瘤样本网分成了 13 社区（如图 4）。此时得到的聚类系数最大，为 Q=0.630（如表 4）。可以看出，肿瘤样本网络的各个社区结构中的节点数目分布也是并不均匀，并且同样存在很多孤立节点。社区内节点间的连接比较紧密，而不同社区间的连接比较稀少。

表 4 肿瘤样本网络不同社区划分的Q值

社区数	10	11	12	13	14	15	16
Q 值	0.565	0.564	0.630	0.629	0.627	0.628	0.626

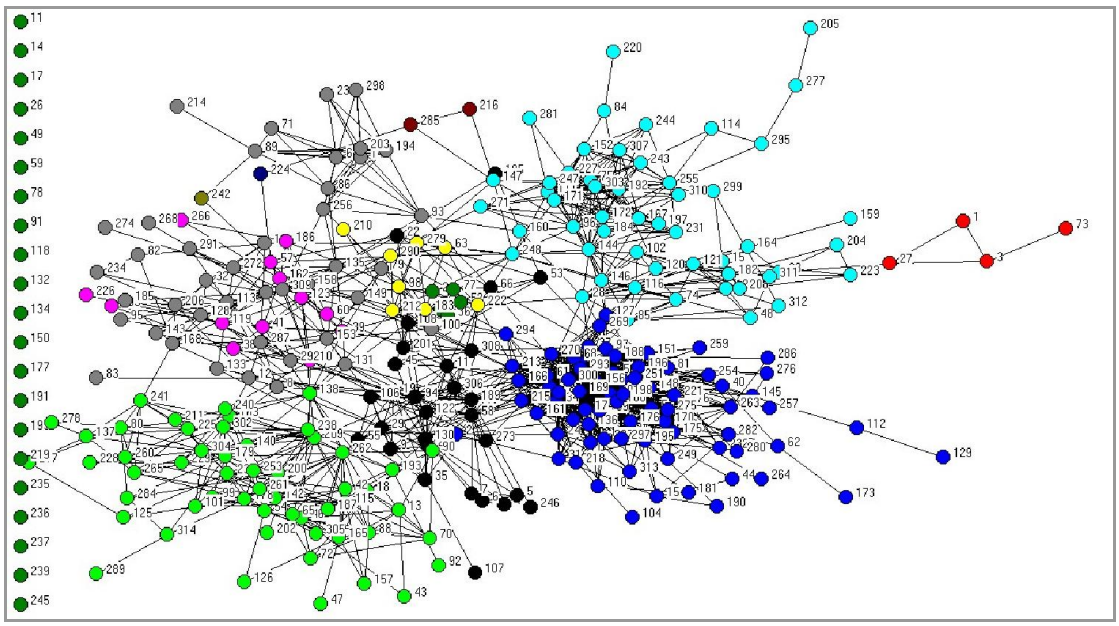

图 4　肿瘤样本网络社区分布图，不同的颜色表示不同的社区

五 结果与讨论

5.1 度序列变化

对于邻接矩阵 MD 与 MN，我们计算出每个节点的度（degree）。我们发现，$\begin{cases} DD_{\max} = 27 \\ DD_{\min} = 0 \end{cases}$，$\begin{cases} DN_{\max} = 47 \\ DN_{\min} = 0 \end{cases}$，其中 $DD_{\max}$、$DN_{\max}$ 分别表示肿瘤样本、正常样本的邻接矩阵中节点的最大度，$DD_{\min}$、$DN_{\min}$ 分别表示肿瘤样本、正常样本的邻接矩阵中节点的最小度。我们发现两个网络中均有度为 0 的节点，说明网络中均有孤立节点，也即有些节点与网络中其他节点只有微弱的相互作用。两个网络中度的最大值发生了明显的变化，说明网络中的有些点与其他点的相互作用强度发生了明显的变化。反应到网络结构中，可以用平均度加以粗略说明，其中肿瘤样本网络的平均度为 9.36，正常样本网络的平均度为 5.28。在肿瘤样本网络中每个基因平均与周围 9.36 个基因有相互作用，在正常样本网中每个基因平均与周围 5.28 个基因有相互作用。

接下来我们来分析度的变化。通过两个网络的度序列做差，我们就能够找到每个节点度的变化情况。下表即为度变化比较大的前十个节点：

表 5 度变化最大的 10 个节点

GenBank Acc No	网络中编号	样本 度值	正常样本度值	度变化值
M94556	61	47	6	41
T70062	156	46	6	40
L28010	34	41	3	38
M37583	169	40	3	37

D14812	87	41	5	36
D29808	300	38	2	36
T65740	139	37	1	36
M22382	68	40	7	33
U30825	56	39	6	33
T84049	215	34	2	32

同时我们对每个节点度的变化值做平均，得到度变化的平均值为 7.0637。其中大于这个平均变化度的节点有 89 个，小于这个平均变化度的节点有 255 个。

我们认为特征基因在这些度变化比较大的节点中的可能性很大。度变化超过平均值的节点与我们查阅的的文献中得出个 34 个特征基因相比对，其中有 15 个基因是它们所共同拥有的（如表 6），我们认为这 15 个基因应该是对我们寻找结肠癌特征基因非常重要的基因。

表 6 度变化最大的节点与文献发现的结肠癌相关基因的交集

GenBank Acc No	网络节点	度变化
M22382	68	33
T96873	180	29
U09564	155	27
H08393	270	24
J02854	213	22
T62947	198	17
M59040	207	17
H20709	2	13
T60155	58	13
T51571	83	11
M36634	248	11
H43887	130	9
X62048	297	9
H64807	296	8
L41559	219	8

5.2 网络比较

两个样本的社区结构图相比较，肿瘤样本图中的边比正常样本中的边明显偏多。实际上有 $\begin{cases} L_{Dis} = 1470 \\ L_{Nor} = 829 \end{cases}$，其中 L_{Dis} 是肿瘤样本中网络的边数，L_{Nor} 为正常样本中网络的边数。这说明在肿瘤发生的过程中，基因之间的相互作用更强了，且影响的“范围”更广了。基因之间相互作用的变化，通过网络中连边的变化都得到了很好的印证。

接下来对我们得到了 15 个重要的基因节点，在网络中分析它们。我们分别在正常样本网络和肿瘤样本网络中找出这些基因节点的位置。观察它们的变化情况。如图 5，图 6 所示

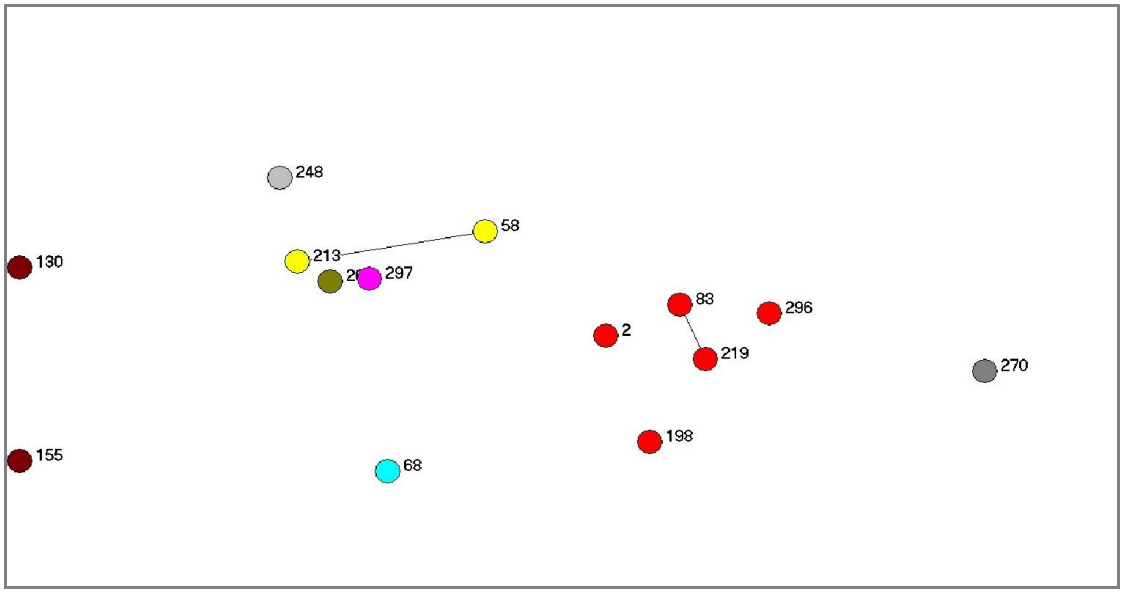

图 5　15 个重要基因在正常样本网络中的分布，不同的颜色表示不同的社区

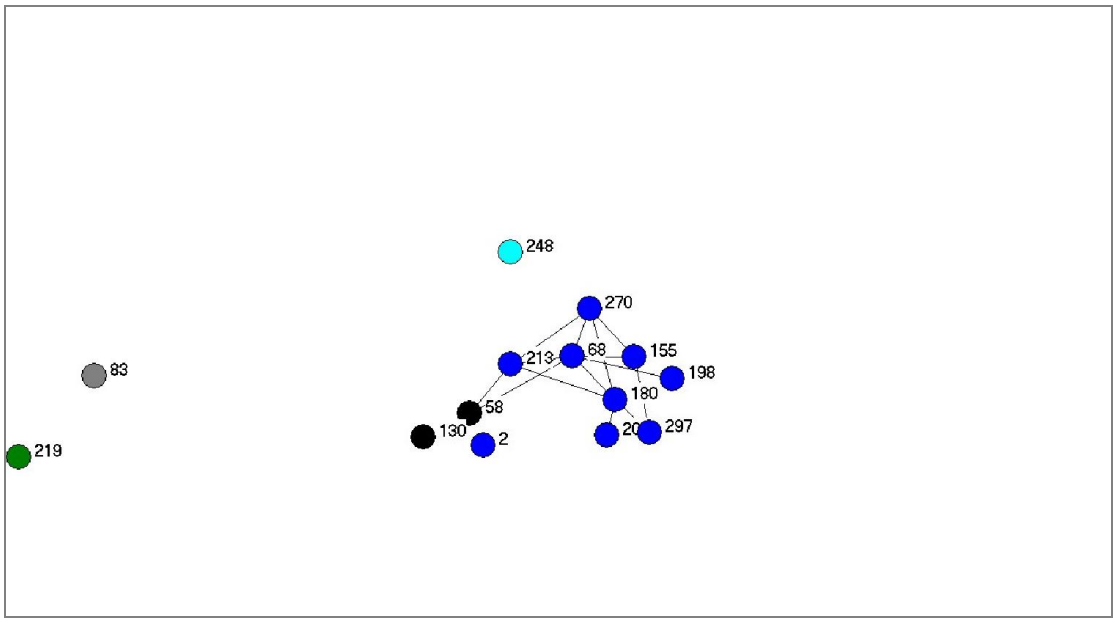

图 6　15 个重要基因在肿瘤样本网络中的分布，不同的颜色表示不同的社区

图 5 为图 2 中仅保留 15 个重要节点的网络图，这 15 个节点在图 2 与图 5 中的相对位置是没有变化的。通过图 5 可以明显的看出 15 个节点分别分布在 7 个不同的社区结构中，其中仅有两对节点相连。通过图 5 我们也并不能发现这些基因节点之间有明显的相互作用。

图 6 为图 4 中仅保留 15 个重要节点的网络图，图 6 中可以明显的看出这 15 个节点的分布就比较集中，其中有 9 个节点都分布在蓝色的社区中，而且这些节点之间的联系也紧密的多。这说明正常样本中的这些基因在发生癌变之后，彼此之间的相互作用明显增强，这对我们提取结肠癌的特征基因具有重要的信息价值。

5.3 网络中的几个特殊节点

在上一步过程中，我们比较了文献中得出的，且度变化较大的 15 个重要节点。发现这 15 个基因在肿瘤特征过程中起了很重要的作用。注意到我们选取的这 15 个基因最大的度变化值是 33，但还有 7 个节点的度变化值超过了 33，却并不在我们查阅的文献的结论中，我们认为有必要在网络中进一步对这些点进行分析。这 7 个基因节点分别是（如表 7）：

表 7

基因编号	GenBank Acc No	网络中编号	度变化值	分类信息指数编号	分类信息指数
440	M94556	61	41	235	0.2247
1067	T70062	156	40	130	0.2651
199	L28010	34	38	158	0.2550
1187	M37583	169	37	206	0.2350
619	H89087	87	36	165	0.2508
1924	H64807	300	36	296	0.2046
973	T65740	139	36	297	0.2045

其中，度变化是同一节点在肿瘤样本网络与正常样本网络中，该节点在两个网络中度的变化值；分类信息指数编号是指该信息指数在所有信息指数中从大到小排列时的次序，我们选取的 314 个基因是分类信息指数 $IIC > 0.2$ 的基因，也即分类信息指数编号前 314 个基因。通过上面的表格我们可以看出，这些基因的分类信息指数都比较大。通常地，样本们会去研究 *IIC* 大的点，分类信息指数编号偏后的那些基因极易在分析的过程中被忽略掉。现在我们发现，这些点在两个网络中度的变化值很大，也即癌变前后这些基因在网络中与其它基因的相互作用有了很大的变化。接下来，我们将这 7 个基因分别放回正常样本和肿瘤样本的网络中去分析它们的变化。如图 7，图 8

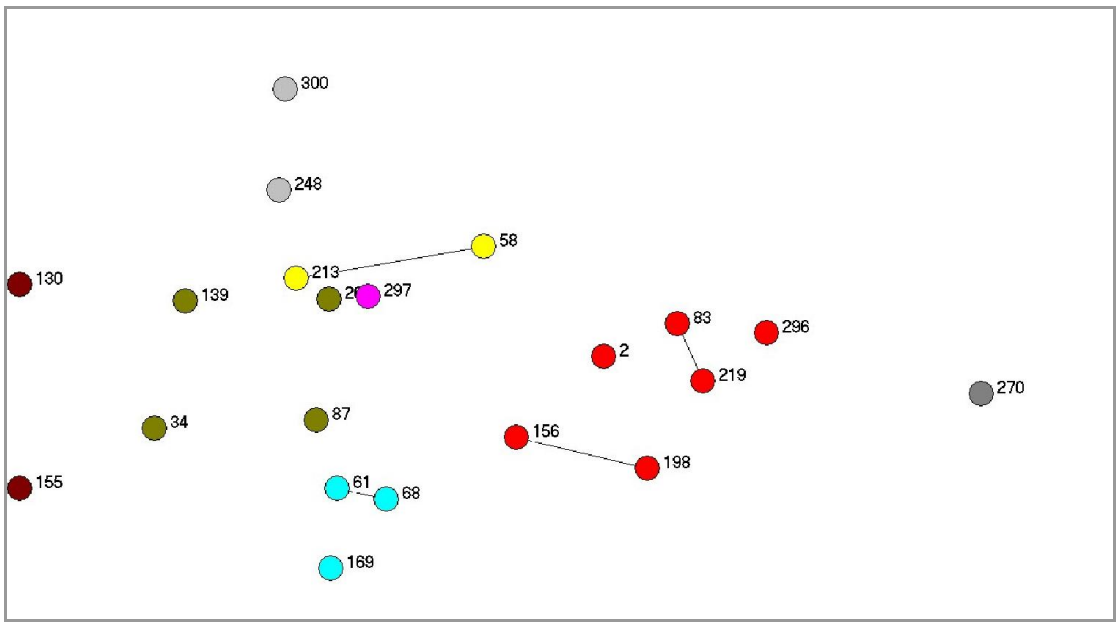

图 7 为我们找到的 15 个重要基因在正常样本中的相对位置。不同的颜色表示不同的社区。同时把度变化最大的 7 个节点（156，87，300，139，169，61，34）也放进了网络中。

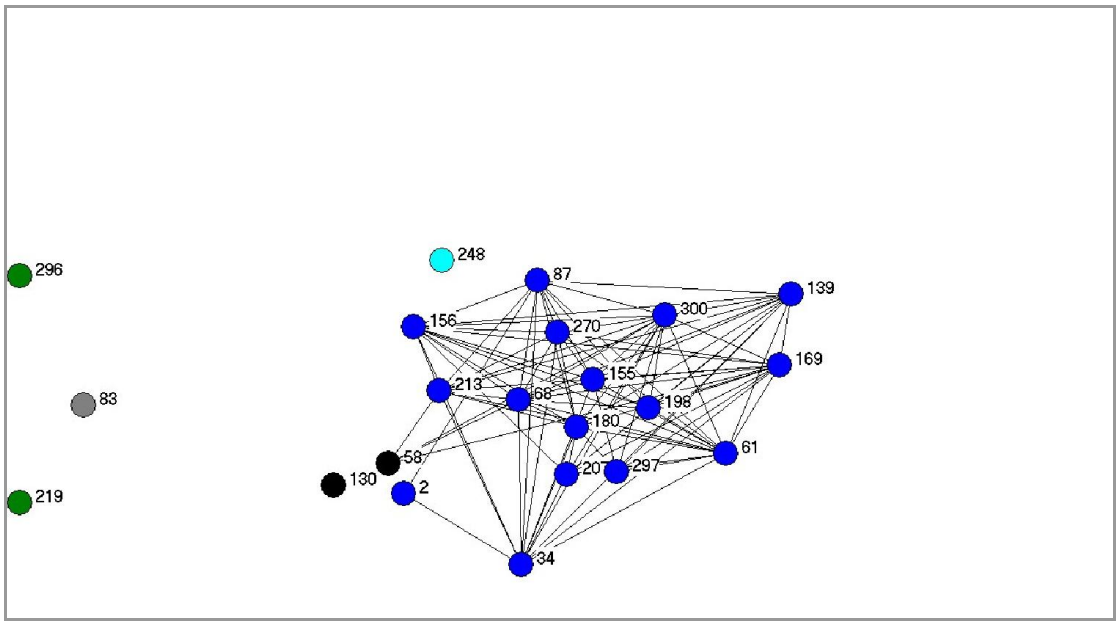

图8为我们找到的15个重要基因在肿瘤样本中的相对位置。不同的颜色表示不同的社区。同时把度变化最大的7个节点（156，87，300，139，169，61，34）也放进了网络中。

从图7中可以观察出，在正常样本网络中，度变化最大的7个节点分别分布在4个社区中，且仅有一个节点与其它节点相连（节点61—节点68）。这说明7个节点在正常样本网络中没有明显的相互作用。而通过观察图 10，我们的发现在肿瘤样本网络中，度变化最大的 7 个节点同时分布在同一个社区中，且这 7 个节点与我们找到的15个重要基因节点中的9个节点（分别为68、180、155、270、213、198、207、2、297）也在同一社区中（图8中蓝色表示的社区），并相连。我们有一个大胆的猜想，结肠癌的特征基因就分布在蓝色所表示的社区中。蓝色社区中的这16个节点所代表的基因分别为M22382,T96873,U09564,H08393,J02854，T62947，M59040，H20709，X62048，及M94556，T70062，L28010，M37583，H89087，H64807，T65740，从功能上看，这些基因对结肠癌的癌变过程发挥了重要的作用。在正常样本网络中，这些点分布的比较分散，而在肿瘤样本网络中，这些点集中到了同一社区中，说明癌变后这些基因之间的相互作用加强。所以这16 个基因就是我们要寻找的结肠癌的特征基因。另外，除了这些在同一社区的节点之外，还有一些散节点落在各个不同的社区中，其中分为两种情况，一种是该基因位于两个社区的连接点处，如节点58（T60155），它是主动脉平滑肌肌动蛋白，而有研究表明肌动蛋白参与DNA转录，所以T60155是我们所寻找的结肠癌的特征基因。另一种是某社区内部的节点，如节点83（T51571），130（H43887），219（L41559），248（M36634），参照这些基因的功能对基因的癌变并没有起到决定性的作用。并且这几个点的度变化值也不是很大，所以，可能是被误选入的，应该被排除掉。综上，本文运用复杂网络的方法，通过社区模块的划分，找出17个结肠癌的特征基因。

六 结论

本文首先通过分类信息指数这一指标对数据做了初步处理，筛选出314个基因节点，剔除了大量的无关基因，对数据进行过滤和降维。并以此分别构建网络

模型。在构建网络模型过程中，我们用两种方法取阈值，得出的结果十分相近，这就说明我们对阈值的选取是合理的。生成网络之后，通过 Newman-Girvan 方法对我们的网络模型划分社区和评价，无论是肿瘤样本网络还是正常样本网络都是很好的社区结构。我们利用度变化值和参考我们查阅文献中得出的结论，挑选出了 22 个基因，其中排除掉 5 个基因后，得出了我们的结论，即结肠癌的特征基因有 17 个。

本文问题研究还有待于进一步加深完善，比如没有考虑到基因筛选后提出的变化不大的点。另外，我们对于生物医学方面的专业知识比较欠缺，在对模块进行分析的时候，对模块的功能分析不够精确。这需要我们以后的继续努力和学习。

[参考文献]：

[1] 程书钧，肿瘤——分子网络病[J]，医学研究杂志，2010,5。

[2] LI Yixin, RUAN Xiaogang, Feature selection for cancer classification based on support vector machine[J]. Journal of Computer Research and Development, 2005,42(10):1796-1801。

[3] GUYON I,WESTON J, BARNILL S, et al. Gene selection for cancer classification using support vector machine[J], Machine Learning, 46(13):389-242. 2000。

[4] 刘全金、李颖新、阮晓钢，基于 SVM 的灵敏度分析方法选取肿瘤特征基因[J]，北京工业大学学报，33(9):954-958.2007。

[5] 刘全金、李颖新、阮晓钢,基于基因表达谱的结肠癌特征基因选取[J], 昆 明理工大学学报(理工版),31(1):89-92,2006。

[6] Xiaosheng Wang and Osamu Gotoh, Microarray-Based Cancer Prediction Using Soft Computing Approach[J], Cancer Informatics, 2009:7,123 - 139。

[7] Xue Wu Zhang, Yee Leng Yap, Dong Wei, Feng Chen1 and Antoine Danchin, Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis[J], European Journal of Human Genetics, 1 - 9,2005。

[8] 李建更、高志坤、严志、阮晓钢,基于双基因分析的结肠癌标志基因选择[J]，中国生物医学工程学报，28(5):691-695,2009。

[9] 张娅 ,饶妮妮 ,王敏 ,徐尚蕾，一种基于基因表达谱的结肠癌特征提取方法[J], 航天医学与医学工程,21(4):356-360,2008。

[10] 王林、戴冠中,复杂网络的 Scale-free 性、Scale-free 现象及其控制[M]，北京,科学出版社,2009。

附录

节点	基因号	节点	基因号	节点	基因号	节点	基因号	节点	基因号	节点	基因号
1	13	30	187	59	419	88	621	117	806	146	1002
2	14	31	189	60	437	89	622	118	807	147	1004
3	16	32	190	61	440	90	625	119	812	148	1014
4	39	33	196	62	444	91	627	120	813	149	1022
5	40	34	199	63	451	92	629	121	817	150	1039
6	41	35	201	64	453	93	634	122	822	151	1042
7	42	36	221	65	489	94	636	123	824	152	1047
8	49	37	241	66	493	95	639	124	826	153	1053
9	59	38	245	67	495	96	648	125	830	154	1058
10	66	39	249	68	513	97	652	126	834	155	1060
11	67	40	264	69	515	98	662	127	866	156	1067
12	70	41	267	70	520	99	663	128	882	157	1073
13	75	42	273	71	523	100	665	129	892	158	1091
14	76	43	281	72	529	101	679	130	897	159	1094
15	78	44	286	73	546	102	689	131	910	160	1108
16	99	45	293	74	548	103	694	132	918	161	1110
17	100	46	295	75	549	104	698	133	929	162	1111
18	102	47	344	76	550	105	732	134	931	163	1115
19	111	48	345	77	559	106	737	135	932	164	1119
20	112	49	350	78	561	107	739	136	936	165	1136
21	125	50	359	79	564	108	758	137	953	166	1139
22	137	51	365	80	570	109	765	138	964	167	1153
23	138	52	370	81	571	110	766	139	973	168	1168
24	141	53	377	82	576	111	779	140	979	169	1187
25	143	54	391	83	581	112	780	141	980	170	1194
26	147	55	395	84	601	113	785	142	982	171	1196
27	166	56	399	85	609	114	792	143	989	172	1197
28	169	57	411	86	617	115	802	144	992	173	1200
29	181	58	415	87	619	116	803	145	994	174	1208

节点	基因号	节点	基因号	节点	基因号	节点	基因号	节点	基因号
175	1209	204	1386	233	1548	262	1698	291	1900
176	1210	205	1387	234	1549	263	1724	292	1904
177	1221	206	1398	235	1559	264	1730	293	1912
178	1223	207	1400	236	1560	265	1758	294	1917
179	1224	208	1401	237	1570	266	1761	295	1920
180	1227	209	1406	238	1573	267	1763	296	1924
181	1231	210	1411	239	1579	268	1770	297	1935
182	1244	211	1414	240	1582	269	1771	298	1938
183	1247	212	1421	241	1583	270	1772	299	1939
184	1248	213	1423	242	1597	271	1785	300	1942
185	1256	214	1439	243	1605	272	1795	301	1943
186	1258	215	1447	244	1608	273	1797	302	1946
187	1260	216	1451	245	1614	274	1801	303	1959
188	1263	217	1452	246	1623	275	1808	304	1960
189	1285	218	1466	247	1634	276	1826	305	1965
190	1286	219	1472	248	1635	277	1836	306	1967
191	1288	220	1473	249	1637	278	1839	307	1972
192	1293	221	1489	250	1644	279	1843	308	1974
193	1325	222	1494	251	1648	280	1867	309	1982
194	1328	223	1495	252	1649	281	1870	310	1983
195	1332	224	1504	253	1651	282	1871	311	1985
196	1334	225	1511	254	1659	283	1873	312	1991
197	1340	226	1514	255	1666	284	1875	313	1993
198	1346	227	1526	256	1668	285	1884	314	1998
199	1348	228	1530	257	1671	286	1886		
200	1360	229	1534	258	1672	287	1887		
201	1365	230	1537	259	1674	288	1892		
202	1366	231	1541	260	1675	289	1894		
203	1381	232	1546	261	1679	290	1897		

注：其中“基因号”表示所选的 314 个基因对应的基因编号； “节点”列表示生成树以及网络中的节点编号。